Construction d'ontologies à partir de textes : la phase de conceptualisation

Thibault Mondary, Sylvie Després, Adeline Nazarenko, Sylvie Szulman

Construction d'ontologies à partir de textes : la phase de conceptualisation

T. Mondary, S. Després, A. Nazarenko, S. Szulman

LIPN - UMR 7030
Université Paris 13 - CNRS
99, avenue J-B Clément - F-93430 Villetaneuse
nom@lipn.univ-paris13.fr

Résumé : Dans cet article nous nous interrogeons sur la manière d'outiller la phase de conceptualisation lors de la construction d'une ontologie à partir de textes. La mise en perspective des résultats obtenus à partir de techniques issues de la terminologie et de la fouille de textes est réalisée selon trois plans (discours, linguistique et conceptuel). Cette étude permet de mieux appréhender les moyens envisageables pour outiller efficacement et de façon cohérente le processus de conceptualisation.

Mots-clés : construction d'ontologies, textes, conceptualisation

La construction d'ontologies à partir de textes constitue un sous-domaine à part entière de l'ingénierie des ontologies. Dans le contexte du Web sémantique, ces ontologies servent essentiellement à l'annotation sémantique de ressources et à la structuration de bases de connaissances. Le recours aux textes est légitimé par les travaux menés en linguistique dont l'hypothèse principale est que les textes sont porteurs de connaissances stabilisées et partagées par des communautés de pratiques. En outre, même s'ils ne les remplacent totalement, les textes sont plus facilement disponibles que les experts qui manquent de temps pour participer au processus de construction. Une ontologie est une spécification formelle d'une conceptualisation d'un domaine, partagée par un groupe de personnes, qui est établie selon un certain point de vue imposé par l'application construite (Studer *et al.*, 1998). Une telle ontologie est constituée d'un ensemble de concepts à la fois organisés hiérarchiquement et structurés par des relations liant ces concepts. Nous ne préjugeons pas ici de l'existence de règles et/ou d'axiomes associés à l'ontologie. Un cadre méthodologique en quatre étapes (constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie) est commun à la plupart des méthodes de construction d'ontologies à partir de textes. Ces étapes, relativement indépendantes, réalisent un double mouvement permettant de passer du niveau textuel (la connaissance est décrite dans des corpus) au niveau conceptuel (la connaissance est décrite *via* des concepts dénotés par les entités linguistiques et les relations entre ces concepts) et de l'informel vers le formel.

Cet article met l'accent sur la phase de conceptualisation qui permet d'articuler le niveau du discours et le niveau ontologique et de penser le passage de l'un à l'autre,

ce qui est essentiel pour toutes les applications où les ontologies doivent servir de base à l'annotation sémantique. Classiquement le passage du niveau du discours au niveau ontologique (Buitelaar *et al.*, 2005) est représenté par un empilement de couches successives qui laisse penser que ce passage se fait de façon séquentielle et dans une même dimension. Or ces différentes couches se situent dans des plans différents : le discours représenté par le corpus, le niveau linguistique constitué d'entités terminologiques et le niveau ontologique constitué des entités de l'ontologie, auxquelles sont éventuellement associées des instances, même si nous ne considérons pas qu'elles fassent partie intégrante de l'ontologie. Distinguer ces trois plans permet de définir de manière plus rigoureuse les entités manipulées (termes *vs.* concepts, relations lexicales *vs.* conceptuelles) et de mieux comprendre le rôle des techniques de traitement automatique des langues (TAL) et de fouille de textes dans le processus de conceptualisation. Cette distinction établie, les étapes préparatoires à ce processus de conceptualisation sont mieux caractérisées (famille de techniques, nature des résultats, etc.) et il devient alors possible d'organiser la cohérence des traitements afin d'outiller ce processus dans le cadre d'une plate-forme dédiée à la construction d'ontologies.

L'article est structuré en trois parties. La première partie présente trois systèmes de construction d'ontologies représentatifs des tendances actuelles où la phase de conceptualisation est réalisée de manière automatique ou semi-automatique. Un exemple issu du corpus du Bureau International du Travail permet d'illustrer les résultats obtenus après les traitements réalisés par ces outils. Dans la seconde partie, nous montrons comment les textes sont exploités pour la conceptualisation en indiquant à la fois quelles sont les informations extraites des textes et comment elles sont utilisées. La discussion de la troisième partie analyse le processus permettant de passer des textes à l'ontologie et pointe quelques verrous qui restent à supprimer pour progresser.

1 Systèmes de construction d'ontologies à partir de textes

Nous avons sélectionné trois systèmes représentatifs des approches évoquées en introduction pour la construction d'ontologies à partir de textes. Nous avons privilégié des systèmes opérationnels, disponibles sur la toile et pouvant exporter au format OWL.

Text2Onto (Cimiano & Volker, 2005) est un outil conçu pour construire des ontologies à partir de textes de manière complètement automatique (voir figure 1). Il est codé en java et est composé de modules qui extraient à partir des textes des concepts[1], des relations entre ces concepts (relation d'équivalence, hiérarchiques, etc.) et des instances de concepts. Chaque module peut utiliser différents algorithmes et combiner leurs résultats : on peut ainsi combiner des patrons d'extraction "à la Hearst" et une ressource comme WordNet pour construire une hiérarchie. Text2Onto utilise l'architecture GATE pour pré-traiter les textes. Les résultats sont dotés d'une mesure de confiance entre 0 et 1 obtenue à l'aide de différentes mesures combinables (TF.IDF, RTF, entropie). De notre point de vue, Text2Onto se présente comme une boîte à outils. L'ontologue doit lui-même sélectionner les algorithmes à utiliser. Il peut accepter ou rejeter

[1] qui s'apparentent selon nous davantage à ce que nous appelons plus loin des candidats-termes.

les résultats obtenus mais pas les modifier ni revenir aux parties des documents dont ils sont issus. Le système KASO (Wang *et al.*, 2006) dont la conception est centrée utilisateur peut être couplé à Text2Onto pour affiner l'ontologie produite à l'aide de méthodes d'acquisition de connaissances telles que la mise en échelle (*laddering*) et le tri par cartes. La nécessité d'avoir recours à des étapes en aval de Text2Onto montre les limites de l'approche tout automatique pour la conceptualisation qui, de notre avis, ne peut se passer de l'intervention humaine.

A partir de notre corpus exemple, Text2Onto extrait 560 "concepts" peu hiérarchisés qui s'apparentent à des mots rarement composés et quelques relations pertinentes (*fire is-a calamity*). Nous avons toutefois choisi de ne pas utiliser de ressource externe (WordNet) pour construire la hiérarchie.

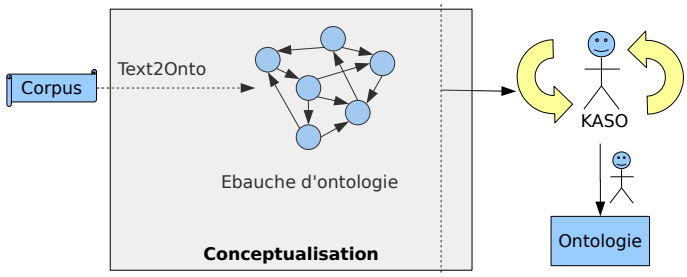

FIG. 1 – Text2Onto + KASO

OntoGen (Fortuna *et al.*, 2006), qui est codé en .net, implémente une approche semi-automatique pour la construction d'ontologies de thèmes (topic ontologies) à partir de collections de documents (voir figure 2). C'est un outil interactif qui suggère à l'expert du domaine des concepts sous la forme de classes de documents, propose une dénotation et leur associe automatiquement des instances (les documents). Il permet de visualiser l'ontologie en cours de construction. OntoGen exploite des algorithmes de fouille de textes non supervisés (k-means(Hartigan & Wong, 1979), LSI(Deerwester *et al.*, 1990)) ou supervisés (svm active learning(Tong & Koller, 2000)) mais toujours selon une approche descendante. A chaque étape le classifieur travaille sur la sous-collection associée au concept qui vient d'être construit. OntoGen propose à l'expert, et c'est à ce dernier de choisir la proposition correcte parmi celles qui lui sont présentées. C'est une approche semi-automatique de la conceptualisation : les outils de classification de documents sont utilisés pour préparer le travail de conceptualisation, l'expert du domaine est guidé dans une démarche descendante mais c'est lui qui construit les concepts et choisit quelles zones de l'ontologie affiner. Sur notre exemple provenant du corpus BIT, OntoGen identifie les concepts du travail forcé (dénoté par les mots clés *compulsory labour, forced compulsory labour*), du travail des enfants (*child, child labour, worst form child labour*), de la liberté syndicale (*workers employment organisation, freedom, associations*) et de la discrimination (*occupation, discrimination, policy*). Il convient toutefois d'insister sur le fait qu'OntoGen se focalise sur la construction d'ontologies de thèmes, et que les instances des concepts sont les documents.

Terminae (Aussenac-Gilles *et al.*, 2008) est une méthode (schématisée sur la figure

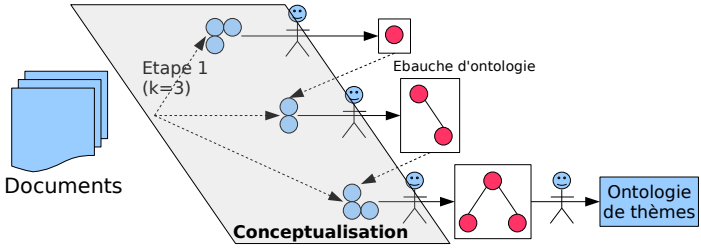

FIG. 2 – OntoGen

3) supportée par un logiciel. Elle propose de guider l'ontologue dans la conception de l'ontologie. Terminae s'appuie sur les résultats des outils de traitement automatique des langues (extracteur de termes, concordancier, détecteur de synonymie, analyseur syntaxique) pour extraire des éléments dans lesquels l'ontologue puise selon ses objectifs de modélisation. Une fiche terminologique affiche les occurrences d'un terme dans le corpus, une ou plusieurs définitions en langage naturel du ou des concepts terminologiques associés et les termes synonymes. Le concept terminologique dénoté par le terme est alors construit par l'ontologue. La phase de conceptualisation, passant par ces fiches, reste entièrement manuelle mais elle est assistée : l'ontologue dispose de différentes vues sur le matériau textuel et Terminae offre une traçabilité entre les concepts de l'ontologie en cours de construction et le corpus (voir figure 3). Cette approche de la conceptualisation n'impose pas de stratégie de construction *a contrario* de la conceptualisation "guidée" d'OntoGen qui impose de fait une stratégie de construction top-down.

Notre corpus exemple traité par l'extracteur de termes Yatea (Aubin & Hamon, 2006) met en exergue des candidats termes plus riches que les "concepts" de Text2Onto, par exemple *abolition of child labour, conditions of employment*. Le terme *forced compulsory labour* retrouvé par OntoGen n'est pas trouvé par Yatea, car il n'existe pas directement dans le corpus mais existe sous la forme *forced **or** compulsory labour*.

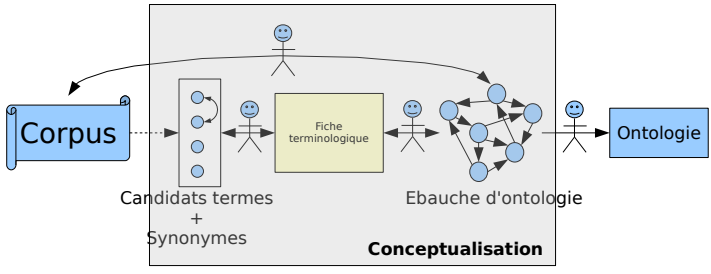

FIG. 3 – Terminae

L'analyse de ces systèmes de construction d'ontologies nous montre la diversité des éléments d'informations qui sont extraits des textes et des techniques utilisées pour

alimenter le processus de conceptualisation.

2 A quoi servent les textes dans la construction d'ontologies ?

Si l'ontologie ne peut être directement extraite des textes, sa construction repose sur les connaissances qui en sont issues. Dans la phase préparatoire de la conceptualisation, on peut distinguer trois opérations principales : l'identification des termes, leur regroupement en classes sémantiques et leur structuration en réseau terminologique. Nous montrons comment chacune de ces opérations alimente le travail de conceptualisation.

2.1 Identifier les termes du domaine

L'analyse terminologique d'un corpus permet d'identifier les syntagmes qui semblent avoir un fonctionnement terminologique, c'est-à-dire qui relèvent d'un vocabulaire spécialisé doté d'une sémantique relativement stable et consensuelle au sein d'une situation de communication déterminée. Diverses méthodes ont été proposées et développées pour extraire des termes de corpus[2] mais le repérage du vocabulaire conceptuel ne peut être entièrement automatisé car les résultats des extracteurs sont bruités et le jugement terminologique est en partie subjectif.

Text2Onto et OntoGen utilisent les mesures statistiques de la recherche d'information pour repérer les termes du corpus[3]. Terminae exploite les sorties d'un extracteur de termes tiers, Lexter (Bourigault, 1995) ou Yatea.

Les termes étant des entités textuelles, ils subissent différentes modifications de surface qui nuisent à leur repérage en corpus et à leur visualisation. Il est donc important, ne serait-ce que pour mesurer leur fréquence et les visualiser, de regrouper les différentes formes d'un même terme (Daille, 2005). Il existe des outils qui permettent de repérer des variantes de termes (cf *supra*) et qui mériteraient d'être intégrés aux systèmes de construction d'ontologie à partir de textes. Identifier la forme canonique commune à un groupe de termes serait également une fonctionnalité précieuse pour guider le choix des étiquettes de concepts.

Une fois effectué ce recensement des candidats termes du corpus, il est important de leur associer un poids et de les trier. Cela permet de focaliser le travail de construction d'ontologie sur un petit nombre de termes et de commencer par les éléments les plus importants[4]. Cette pondération peut exploiter des critères variés. Certaines mesures reposent sur la structure du vocabulaire du corpus (la fréquence des termes notamment), sur la composition des termes, leur degré de spécialisation (Drouin, 2006) ou leur saillance en corpus en tenant compte des marques d'emphase, de la structure logique des documents, etc. (Mekki & Nazarenko, 2004).

[2] Voir Jacquemin & Bourigault (2003) pour une revue d'ensemble des outils d'analyse terminologique.

[3] (TF.IDF complété de RTF et d'entropie pour Text2Onto).

[4] L'extraction terminologique produit généralement des listes importantes de candidats-termes (typiquement, plusieurs milliers voire quelques dizaines de milliers de candidats pour un corpus de 100 000 mots).

Même s'ils en sont souvent le reflet, les termes ne sont pas des concepts et passer des termes aux concepts relève de la tâche de conceptualisation. Le fait d'exploiter des extracteurs de termes, de regrouper les variantes de termes proches et de les trier par ordre de pertinence guide le processus de conceptualisation sans toutefois l'automatiser. La démarche préconisée dans la méthode Terminae consiste à sélectionner les termes jugés centraux du domaine et à commencer à construire l'ontologie à partir de ceux-ci mais Terminae laisse l'ontologue libre d'appréhender la conceptualisation selon ses désirs, de manière descendante par exemple.

2.2 Construire des classes sémantiques

Une autre approche consiste à regrouper les mots en classes sémantiques susceptibles de représenter les concepts du domaine. Cette approche repose sur l'hypothèse harrissienne selon laquelle le sens des mots peut se déduire de leurs emplois (Harris *et al.*, 1989). De nombreuses variantes de cette méthode ont été proposées (Lin & Pantel, 2002). Elles diffèrent par la définition du contexte qu'elles prennent en compte (fenêtre de mots *vs.* dépendances syntaxiques), par la mesure de similarité qu'elles utilisent et par la méthode d'apprentissage utilisée. La méthode LSI repose sur la même démarche distributionnelle mais elle est dépendante de la taille des contextes et cause parfois des difficultés en cas de proximité du vocabulaire utilisé. Elle peut être améliorée par l'ajout de connaissances syntaxiques (Roche & Chauché, 2006) et de connaissances sémantiques (N.Béchet *et al.*, 2008). OntoGen intègre les algorithmes k-means et LSI pour proposer à l'utilisateur des candidats-concepts. Text2Onto et Terminae n'exploitent pas cette approche.

L'utilisation des classes sémantiques pour le travail de conceptualisation ne va cependant pas de soi. On ne peut pas traduire directement une classe en concept. Les classes obtenues ont une certaine cohérence sémantique mais il faut généralement les nommer, en retoucher les contours, les scinder, etc. La démarche proposée dans ASIUM (Faure & Nedellec, 1999) est intéressante de ce point de vue parce qu'elle repose sur un processus coopératif de classification hiérarchique ascendante. Une autre difficulté vient de ce que cette approche par classification est difficile à combiner avec l'approche terminologique précédente qui ne repose pas sur les mêmes unités textuelles.

2.3 Ebaucher la structuration

Les textes expriment des informations sur les relations sémantiques que les termes entretiennent entre eux. Si l'on considère que les termes représentent les concepts de l'ontologie à construire, les relations sémantiques qu'ils entretiennent peuvent être considérées comme le reflet de relations conceptuelles et repérer ces relations aide à structurer l'ontologie. Différents types de relations sémantiques sont exploités pour la construction d'ontologies : relations hiérarchiques comme dans les thésaurus (hyperonymie ou méréonymie) ou relations plus spécialisées (par ex. *avoir pour indication* en médecine).

L'extraction des relations terminologiques est cependant une tâche complexe du fait de la diversité des relations sémantiques à prendre en compte et de la diversité des méthodes mises en oeuvre. On trouve des méthodes structurelles exploitant la struc-

ture interne des termes et des méthodes contextuelles qui reposent sur le contexte d'emploi des termes. Les méthodes utilisées pour le regroupement sémantique peuvent également servir à la structuration. Certaines proposent une organisation des classes en hiérarchies. Dans la famille des méthodes de regroupement non supervisées, on distingue les méthodes agglomératives (plus proche voisin, distance maximum...) qui regroupent des clusters existants selon des mesures de similarité, des méthodes divisives (bisection k-means). L'analyse formelle de concepts telle qu'utilisée par (Cimiano *et al.*, 2005) identifie tous les groupes de mots ayant des contextes syntaxiques communs (dits concepts), la hiérarchisation des groupes créés (treillis de concepts) et la factorisation maximale des propriétés entre ces groupes. Malheureusement, il n'existe pas d'outil combinant ces différentes méthodes qui gagneraient à l'être car elles ne sont pas toutes également fiables et productives sur tous les corpus. Text2Onto permet de repérer des relations hiérarchiques à l'aide d'approches structurelles, contextuelles et de ressources externes (WordNet). Des relations d'équivalence entre concepts sont proposées à partir d'une analyse contextuelle. OntoGen repère des clusters de termes, et c'est l'utilisateur expert du domaine qui structure l'ontologie à l'aide des propositions du logiciel. Dans le cas de Terminae, la structuration de l'ontologie est essentiellement manuelle même si Terminae exploite les résultats de SynoTerm, qui extrait des relations de synonymie entre termes (Hamon *et al.*, 1998).

3 Discussion

3.1 Penser le passage des textes à l'ontologie

Construire une ontologie à partir de textes nécessite d'opérer un double changement de perspective. Il faut d'abord passer du linguistique au conceptuel. Certes, l'ontologie peut être utilisée comme modèle de connaissance pour interpréter des textes et la langue est un moyen privilégié d'expression de la connaissance, mais les deux niveaux linguistique et conceptuel ne sont pas le reflet direct l'un de l'autre. Passer du discours, pris comme réalisation linguistique avec ses répétitions, son développement particulier, un ancrage situationnel spécifique, à un modèle constitue un second changement de perspective, qui demande lui-aussi à être clarifié. Nous considérons que trois plans différents sont en jeu dans le processus de conceptualisation : le plan du discours (le corpus), le modèle linguistique et le modèle conceptuel (l'ontologie et ses instances). Il est essentiel de montrer comment ils s'articulent l'un à l'autre pour analyser et définir précisément le rôle de ce processus.

Passer des textes à une ontologie, c'est donc passer de la réalisation linguistique au modèle conceptuel. Le modèle linguistique joue un rôle charnière. On trouve à ce niveau la description des éléments de vocabulaire (dictionnaire d'entités nommées, terminologie, thésaurus, classes sémantiques, etc.) associée à certains éléments de grammaire (règles contextuelles de désambiguïsation). Ce qui est représenté à ce niveau, c'est le sous-langage représenté par le corpus et non la langue dans toute sa généralité. Ce modèle de la langue est une construction abstraite, obtenue par induction à partir des réalisations observées en corpus en exploitant les résultats des outils de traitement automatique des langues et de fouille de texte mentionnés plus haut.

Une approche classique de construction d'ontologie repose sur l'idée qu'il est possible d'extraire des terminologies des corpus et que les termes ainsi mis en évidence sont le reflet des concepts du domaine. On établit ainsi une double correspondance entre les unités textuelles et les unités terminologiques de la langue d'une part, puis entre ces dernières et les concepts. Aucune de ces deux correspondances n'est cependant bijective et dans les deux cas une intervention humaine est nécessaire : 1/ A une unité terminologique correspond plusieurs occurrences textuelles et à l'inverse, les unités textuelles ambiguës peuvent être associées à plusieurs termes. 2/ De la même manière, les termes retenus au niveau de l'analyse terminologique peuvent avoir 0, 1 ou plusieurs correspondant conceptuels et certains concepts n'ont pas de pendant terminologique.

On pourrait considérer que la démarche consistant à construire des concepts extensionnellement à partir des noms d'entités nommées –plus couramment appelées "noms propres"– repérées en corpus et dont on fait l'hypothèse qu'elles renvoient à des instances de concepts (Cimiano, 2006) n'utilise pas de modèle linguistique. Cependant, très souvent, on exploite un dictionnaire d'entités nommées qui joue plus ou moins le même rôle que la terminologie. Il liste toutes les formes sous lesquelles une entité nommée peut être exprimée en corpus et il leur associe un type sémantique qui va guider son rattachement à un concept de l'ontologie.

Construire une ontologie à partir de corpus ne peut donc être fait de manière entièrement automatique. La langue reflète le modèle du domaine mais de manière partielle et déformée[5]. Des choix de modélisation sont à faire. Les textes ne désignent pas les concepts du domaine[6] même si, à travers le vocabulaire employé, on voit poindre des notions importantes, même si on peut se faire une idée de l'envergure du domaine à modéliser et donc de la couverture de l'ontologie à produire, il reste un important travail de conceptualisation à faire. Il concerne au premier chef la construction des concepts : a/ faut-il regrouper différents termes sous un même concept ou au contraire introduire des distinctions conceptuelles non reflétées dans la terminologie d'une langue ? b/ faut-il insérer un noeud de structuration sans équivalent terminologique dans l'ontologie (Aussenac-Gilles *et al.*, 2005) pour regrouper des concepts apparentés ? mais aussi et plus fondamentalement, la forme sous laquelle sont représentées les notions dans l'ontologie : a/ est-il préférable de représenter la notion de "travail obligatoire" par des concepts ou par des relations ? b/ faut-il considérer le "Bureau International du Travail" comme une instance ou comme un concept ?

Les réponses apportées à l'ensemble de ces questions tiennent à la granularité de la description ontologique à produire, mais aussi aux inférences que l'ontologie doit pouvoir supporter à son exploitation.

---

[5] Cette déformation tient à la fois à la logique argumentative qui impose de passer sous silence les connaissances partagées, aux choix stylistiques de l'auteur (raccourcis, métaphores) et à l'économie propre de la langue qui contraint la mise en mots.

[6] A noter que l'opération inverse consistant à annoter des textes à partir d'une ontologie n'est pas immédiate non plus. Elle nécessite des règles de désambiguïsation pour passer du concept au terme et du terme aux occurrences.

3.2 Supprimer des verrous

Si la démarche présentée ci-dessus paraît viable pour construire une ébauche d'ontologie en support au travail de modélisation, il nous semble cependant que certains verrous restent à supprimer pour avoir des plates-formes cohérentes et complètes de construction d'ontologies à partir de textes.

3.2.1 Exploiter à la fois les termes et les entités nommées

Le premier verrou concerne le rôle respectif des termes et des entités nommées dans la construction d'ontologie. Le problème est à la fois linguistique et technique.

L'opposition que fait la linguistique entre les termes (unités linguistiques relevant d'un vocabulaire spécialisé) et entités nommées est souvent exploitée pour désigner respectivement des concepts et des instances. On utilise donc les extracteurs de termes pour repérer des concepts et l'extraction des entités nommées pour peupler l'ontologie. Cette démarche repose sur l'hypothèse que les noms propres sont des noms d'entités qui désignent de manière univoque des entités référentielles. Or certaines entités nommées sont ambiguës et certaines expressions définies désignent des entités référentielles aussi clairement que des noms propres. La démarche associant les termes à des concepts et les entités nommées à des instances a donc ses limites. Le problème se complique du fait que les méthodes utilisées pour retrouver les termes et les entités nommées dans un corpus sont elles-mêmes différentes. L'analyse terminologique repose sur des critères de composition morphosyntaxique et de récurrence tandis que l'extraction des entités nommées repose davantage sur des règles typographiques et contextuelles. Quand on les applique sur le même texte, on obtient deux ensembles d'unités textuelles plus ou moins disjoints, chacun comportant sa part de silence et de bruit. Toute la question consiste alors à bien définir le rôle de chaque méthode et de définir une méthodologie de conceptualisation qui exploite au mieux les deux.

3.2.2 Articuler les approches symboliques et distributionnelles

Si le travail de conceptualisation ne peut être automatisé, il peut être assisté. On l'a vu plus haut, il peut exploiter de multiples éléments extraits des textes mais ces éléments sont produits par des outils différents et ils sont difficiles à combiner.

Une difficulté importante à laquelle on est confronté pour construire des ontologies à partir de textes est la diversité des outils d'analyse de corpus. Prenons l'exemple de la structuration de l'ontologie en hiérarchie de concepts. Plusieurs méthodes sont proposées. L'approche à partir de patrons explore le corpus à la recherche de mentions explicites de liens d'hyperonymie entre termes ("X est une sorte de Y") et se propose de les interpréter en lien de subsomption entre concepts ($Cx < Cy$). L'approche distributionnelle fait l'hypothèse que le sens des mots se reflète dans la manière dont ils sont employés, c'est-à-dire dans leur distribution, et s'en sert pour construire des classes de mots qui ont des distributions proches. Les méthodes de classification hiérarchiques permettent alors de hiérarchiser les classes sémantiques obtenues et on fait l'hypothèse qu'à chaque classe correspond un concept. Ces différentes approches n'ont pas du tout la même portée : la première fournit généralement peu de liens de subsomption mais

est assez fiable ; la seconde est à la fois plus productive et plus bruitée car les distributions captent aussi bien des associations conceptuelles que des tournures linguistiques ; la troisième est complémentaire de la seconde.

Articuler ces différentes approches pour permettre au travail de conceptualisation d'exploiter l'ensemble du matériau textuel de manière cohérente relève encore de la gageure. Il faudrait pour cela fournir une boîte à outils à l'utilisateur en l'assortissant de recommandations méthodologiques mais aussi présenter l'ensemble des résultats dans une interface unique, en les typant ou en les pondérant différemment selon leur origine. Il faudrait également assister le travail de validation par des outils de vérification de cohérence. Cela suppose en amont que les résultats produits par les différents outils soient comparables à défaut d'être identiques ou compatibles. Or, les approches distributionnelles travaillent aujourd'hui sur les mots plutôt que sur les termes qui sont souvent des unités polylexicales. Coupler un outil de classification distributionnelle de mots et un extracteur de termes n'a donc pas de sens. Il est urgent à notre sens de résoudre ce type de problème pour avoir des plates-formes de construction d'ontologies exploitables.

3.2.3 Acquérir des relations

L'acquisition des relations constitue sans doute le principal verrou aujourd'hui. On retrouve comme dans le cas de la structuration hiérarchique l'inconvénient d'avoir des méthodes hétérogènes pour l'acquisition des relations (méthodes à base de patrons, méthodes reposant sur la composition syntaxique ou sémantique des termes, approche distributionnelle pour pondérer les relations découvertes, etc.). Là aussi il est important de tirer les leçons du passé pour bien comprendre comment tirer profit de chaque méthode et les combiner au mieux. Le second problème vient de ce que la démarche d'acquisition de relation est très dépendante à la fois du type de relation à acquérir et du genre des corpus. Cela revient à dire que l'extraction de relations est davantage guidée par l'ontologie que l'extraction des termes, qui repose sur des méthodes plus génériques. Cela signifie aussi que l'acquisition de relations doit reposer sur un processus coopératif où l'utilisateur forge ses outils de fouille de textes en fonction des informations qu'il recherche. Peu de travaux vont dans ce sens.

4 Conclusion

La phase de conceptualisation est une phase cruciale de la construction d'une ontologie à partir de textes qui effectue le passage du niveau du discours au niveau ontologique. Dans ce papier, nous revenons sur la présentation habituelle en couches successives de cette articulation et nous proposons de distinguer trois plans (le discours, le linguistique et l'ontologique, les instances de l'ontologie constituant la frontière entre l'ontologie et le corpus) afin de mieux appréhender les difficultés inhérentes au passage dans chacun de ces plans. Une fois cette distinction établie, les entités manipulées sont définies de façon plus rigoureuses et le rôle des techniques de TAL et de fouille de textes est plus facilement analysé. En outre, il devient possible de s'interroger sur l'organisation des traitements préparatoires au processus de conceptualisation et de préparer

ainsi des réponses aux questions soulevées par le développement de plates-formes de construction d'ontologies telle que celle envisagée pour le projet DaFOE4App[7]. A l'issue de ce tour d'horizon, il semble que la réflexion peut s'orienter vers des méthodes hybrides alliant les résultats de méthodes terminologiques à ceux issus de la fouille de textes, à condition de bien maîtriser les caractéristiques et les performances de chacune d'entre elles.

Références

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In T. SALAKOSKI, F. GINTER, S. PYYSALO & T. PAHIKKALA, Eds., *Advances in Natural Language Processing (Proceedings of the 5th International Conference on NLP (FinTAL'06*, LNAI 4139, p. 380–387 : Springer.

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2005). *Modélisation du domaine par une méthode fondée sur l'analyse de corpus*, In *Ingénierie des connaissances*, p. 49–71. L Hamattan.

AUSSENAC-GILLES N., DESPRES S. & SZULMAN S. (2008). *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*, chapter The Terminae Method and Platform for Ontology Engineering from Texts, p. A paraitre. IOS Press.

BOURIGAULT D. (1995). Lexter, a terminology extraction software for knowledge acquisition from texts. In *9th Banff Knowledge Acquisition for knowledge Based Systems Workshop*, Banff.

P. BUITELAAR, P. CIMIANO & B. MAGNINI, Eds. (2005). *Ontology Learning from Text : Methods, Evaluation and Applications*, volume 123 of *Frontiers in Artificial Intelligence*. IOS Press.

CIMIANO P. (2006). *Ontology Learning and Population from Text : Algorithms, Evaluation and Applications*. Springer.

CIMIANO P., HOTHO A. & STAAB S. (2005). Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research (JAIR)*, **24**, 305–339.

CIMIANO P. & VOLKER J. (2005). Text2onto - a framework for ontology learning and data-driven change discovery. In A. MONTOYO, R. MUNOZ & E. METAIS, Eds., *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, p. 227–238, Alicante, Spain : Springer.

DAILLE B. (2005). Variations and application-oriented terminology engineering. *Terminology*, **11**(1), 181–197.

DEERWESTER S. C., DUMAIS S. T., LANDAUER T. K., FURNAS G. W. & HARSHMAN R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6), 391–407.

DROUIN P. (2006). Termhood experiments : quantifying the relevance of candidate terms. In H. PICHT, Ed., *Modern Approaches to Terminological Theories and Applications*, volume 36 of *Linguistic Insights*, p. 375–391. Dordrecht, NL : Peter Lang AG.

[7] `http://www.dafoe4app.fr`

FAURE D. & NEDELLEC C. (1999). Knowledge acquisition of predicate argument structures from technical texts using machine learning : the system asium. In D. F. ET R. STUDE, Ed., *Proceedings of the 11th International Conference on Knowledge Engineering and Knowledge Management*, p. 329–334 : Springer-Verlag.

FORTUNA B., GROBELNIK M. & MLADENIC D. (2006). Semi-automatic data driven ontology construction system. In *Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia*.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *Proceedings of 17th International Conference on Computational Linguistics (COLING-ACL'98)*, p. 498–504, Université de Montréal, Montréal, Quebec, Canada.

HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK J. P., DALADIER A., HARRIS T. & HARRIS S. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*, volume 104 of *Boston Studies in the Philosophy of Science*. Boston : Kluwer Academic Publisher.

HARTIGAN J. A. & WONG M. A. (1979). Algorithm AS 136 : A k-means clustering algorithm. *Applied Statistics*, **28**(1), 100–108.

JACQUEMIN C. & BOURIGAULT D. (2003). Term extraction and automatic indexing. In R. MITKOV, Ed., *Handbook of Computational Linguistics*, chapter 19, p. 599–615. Oxford, GB : Oxford University press.

LIN D. & PANTEL P. (2002). Concept discovery from text. In *Proceedings of Conference on Computational Linguistics (COLING-02)*, Lecture Notes in Computer Science , Vol. 4011, p. 577–583.

MEKKI T. A. E. & NAZARENKO A. (2004). Une mesure de pertinence pour le tri de l'information dans un index de fin de livre. In *Actes de la Conférence Internationale sur le Traitement Automatique des Langues Naturelles (TALN'04)*, Fès, Maroc.

N.BÉCHET, ROCHE M. & CHAUCHÉ J. (2008). Utilisation d'informations syntaxico-sémantiques associées à lsa pour améliorer les méthodes de classification conceptuelles. In *Dans les actes de la conférence EGC'2008, 29 janvier–1er février 2008, Sophia-Antipolis, France*, p. 589–600.

ROCHE M. & CHAUCHÉ J. (2006). Lsa : les limites d'une approche statistique. In *Dans les actes de l'atelier FDC'06 (Fouille de Données Complexes dans un processus d'extraction des connaissances) à la conférence EGC'2006, 17-20 janvier 2006, Villeneuve d'Ascq, France*, p. 95–106.

STUDER R., BENJAMINS V. R. & FENSEL D. (1998). *Knowledge Engineering : Principles and Methods*, In *Data Knowl. Eng.*, volume 25, p. 161–197.

TONG S. & KOLLER D. (2000). Support vector machine active learning with applications to text classification. In P. LANGLEY, Ed., *Proceedings of ICML-00, 17th International Conference on Machine Learning*, p. 999–1006, Stanford, US : Morgan Kaufmann Publishers, San Francisco, US.

WANG Y., VOLKER J. & HAASE P. (2006). Towards semi-automatic ontology building supported by large-scale knowledge acquisition. In *AAAI Fall Symposium On Semantic Web for Collaborative Knowledge Acquisition*, volume FS-06-06, p. 70–77, Arlington, VA, USA : AAAI AAAI Press.